U.S. Forest Service

RESEARCH NOTE

CENTRAL STATES FOREST EXPERIMENT STATION
COLUMBUS, OHIO — R. D. LANE, DIRECTOR

INTERMOUNTAIN STATION
Central Reference File
No. 073

CS-2 March 1963

THE LEAST SQUARES FIT OF AN ALGEBRAICALLY UNSPECIFIED FORM

There are occasions when an accurate yet simple algebraic description of an established graphic form cannot readily be devised. Under such circumstances, the graphic form itself may be used as an independent variate transform to be scaled to any data-set Y_i by Least Squares.

Consider the case where a sensibly smooth curve has been fitted by "eye"[1/] through a set of plotted data points and it is desired to fit this curve through the same data or new data by the Least Squares Fitting System. One may proceed as follows:

Let a two-dimensional, graphed curve be based on, say, 30 paired values of Y and X, dependent and independent variates respectively. For each of the 30 X-values, read the corresponding graphic-curve value of Y. Call these values X_T and let them take the place of the usual algebraic transform(s) of X, to be scaled to the Y_i by Least Squares. Then, fit the model

$$Y = a + bX_T$$

through the Y_i using the 30 paired values of Y and X_T, where $\hat{b}$[2/] will be close to 1.00 if the graphic fit approaches the Least Squares fit, and $\hat{a}$ will be close to zero.

The total sum of squares of deviations of Y_i from $\bar{Y}$ is

$$SS_{total} = \Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{n}$$

The sum of squares attributable to regression is

$$SS_{reg.} = \hat{b}\,\Sigma xy \qquad \text{where: } \hat{b} = \Sigma xy/\Sigma x^2$$

$$\Sigma xy = \Sigma X_T Y - \Sigma X_T \Sigma Y/n$$

$$\Sigma x^2 = \Sigma X_T^2 - (\Sigma X_T)^2/n$$

CENTRAL FILE COPY

1/ i.e.: An approximate Least Deviations Fitting System.
2/ ^ denotes an estimated quantity.

And, the sum of squared deviations about the form which is fit through the data is

$$SS_{res.} = SS_{total} - SS_{reg.}$$

$$R^2 = SS_{reg.}/SS_{total}$$

A test of significance for regression would be

$$F_{1,n-2} = \frac{SS_{reg./1}}{SS_{res./n-2}} = \frac{MS_{reg.}}{MS_{error}}$$

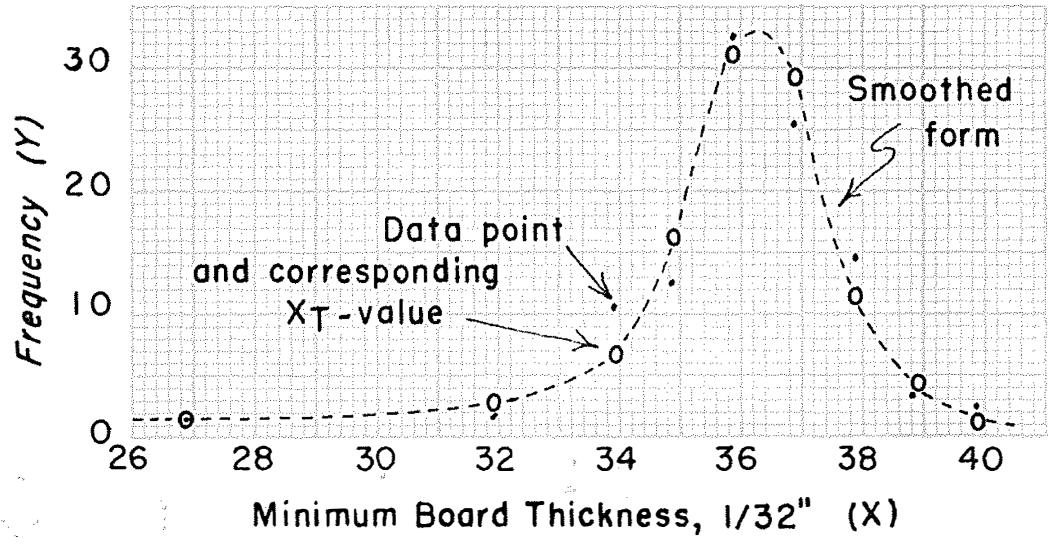
but is subject to question when the X_T-form is not derived independently of the data through which it is being fit by Least Squares.

EXAMPLE

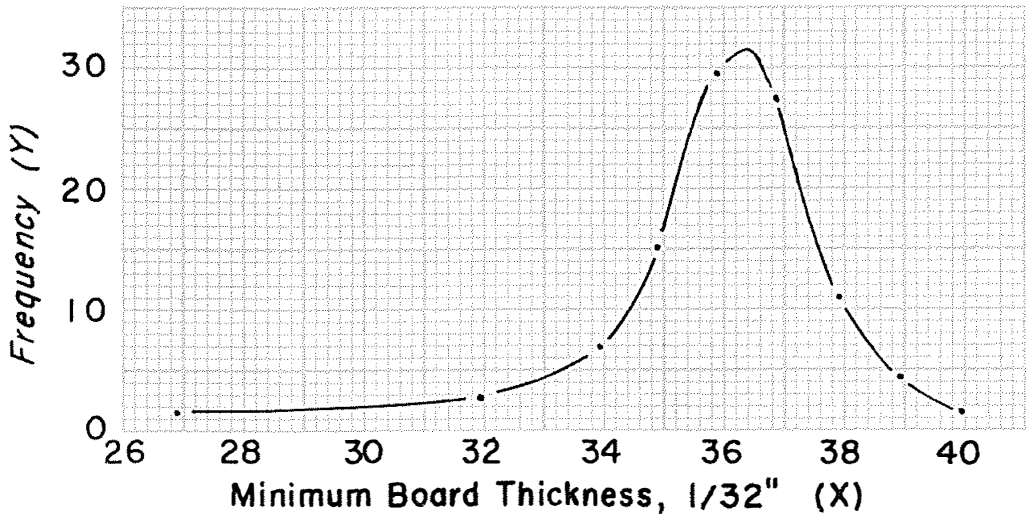
A sawmill study of 100 boards provided an array of frequencies by board-thickness groups, to which the fitting method described herein was applied--see the graphs and computations that follow.

Original Data		Smoothed Form	Least Squares Fit of Smoothed Form
X	Y	X_T	$\hat{Y} = .788 + .9272\ X_T$
27	1	.7	1.4 = .788 + .649
32	1	2.1	2.7 = .788 + 1.947
34	10	6.2	6.5 = .788 + 5.749
35	12	15.2	14.9 = .788 + 14.093
36	32	31.0	29.5 = .788 + 28.743
37	25	29.0	27.7 = .788 + 26.890
38	14	11.0	11.0 = .788 + 10.199
39	3	4.0	4.5 = .788 + 3.709
40	2	1.0	1.7 = .788 + .927
Σ	100	100.2	

SMOOTHED FORM THROUGH PLOTTED (X,Y) VALUES AND ASSOCIATED X_T VALUES



LEAST SQUARES FIT OF ORIGINAL, SMOOTHED FORM TO THE Y_i --- $\hat{Y} = .788 + .9272\ X_T$



$$\Sigma X_T^2 \text{ ---------- } = 2{,}214.38$$

$$(\Sigma X_T)^2/n = (100.2)^2/9 = 1{,}115.56$$

$$\Sigma x^2 \text{ ---------- } = 1{,}098.82$$

$$\Sigma X_T Y \text{ ---------- } = 2{,}132.20$$

$$\Sigma X_T \Sigma Y/n = (100.2)(100)/9 = 1{,}113.33$$

$$\Sigma xy \text{ ---------- } = 1{,}018.87$$

$$\hat{b} = \Sigma xy/\Sigma x^2 = 1{,}018.87/1{,}098.82 = .92724$$

$$\hat{a} = \bar{Y} - \hat{b}(\bar{X}_T) = 100/9 - .92724(11.1333) = .788$$

Then--- $\hat{Y} = .788 + .9272\ X_T$

And--- $\Sigma Y^2 \text{ ---------- } = 2{,}104.0$

$$(\Sigma Y)^2/n = (100)^2/9 = 1{,}111.1$$

$$\Sigma y^2 \text{ ---------- } = 992.9$$

$$SS_{total} = \Sigma y^2 = 992.9$$

$$SS_{reg.} = \hat{b}\ \Sigma xy = .92724\ (1{,}018.87) = 944.7$$

$$SS_{res.} = 992.9 - 944.7 = 48.2$$

$$R^2 = 944.7/992.9 = .95$$

Had the curve form adopted been developed from a prior set of related data, or from hypothesis, tests of regression significance would be appropriate, as follows:

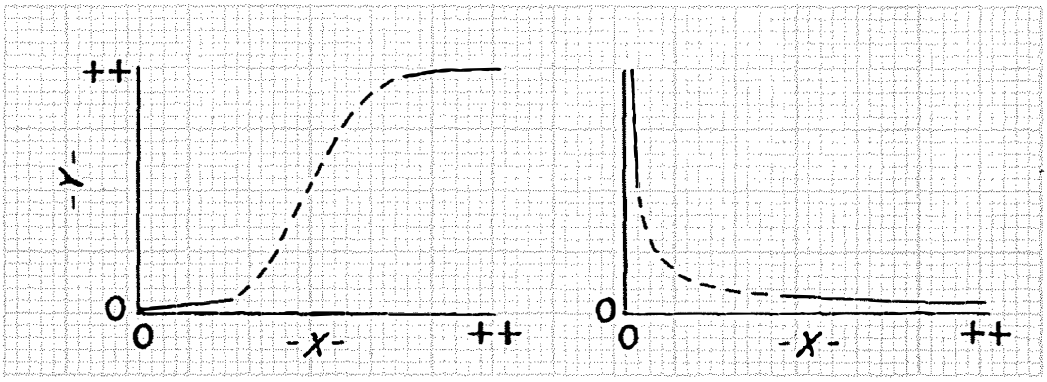
$F_{1,(9-2)} = \frac{944.7}{48.277} = 137$ -- which, of course, is significant beyond the 1 percent level.

To obtain $\hat{Y}$ for other than the original 30 X-values, either return to the original graph for corresponding X_T and solve for $\hat{Y}$ in the derived form $\hat{Y} = \hat{a} + \hat{b}X_T$; or perhaps more simply, plot $\hat{Y}$ for, say, 10 values of X across the range of X. Then join such points by straight lines or suitably smoothed curves, and read $\hat{Y}_i$ directly from the graph.

To fit the original, smoothed form to a new set of (X,Y)-values, simply read the X_T for the new X-values and proceed as above.

The fitting system is analogous for cases involving more than two dimensions and permits the contribution of one or more variates to $SS_{reg.}$ in any fitting order to be evaluated when alternative regressions are graphed and then fitted as above.

EXTREME CAUTION is urged in developing the "sensibly smoothed curve," which is assumed for the fitting process described here. Before looking at your data, construct an expected response form covering the extremes of the independent variate. Usually you can be quite certain as to the slope and value of the curve at such extremes, e.g. flat, tending to zero--or vertical, tending to infinity. Then, what happens to the curve between the extremes becomes self-evident--see the dotted lines in the graphed examples below:



Points in the range of the independent variate at which important changes in slope occur are sometimes predictable from past experience. Exert every effort to bring known information to bear on this estimate--even quantification where possible. Then observe the shape of the expected curve in the range of the independent variate relevant to your data. Fit a similar curve through your data points, allowing the data to guide you as to slope and points of critical slope change. The resulting form constitutes a "sensibly smoothed curve" for the purposes of this paper.

Chester E. Jensen, statistician
Central States Forest Experiment Station